2013

Authorship Attribution on the Enron Email Corpus

Xuan Li

AUTHORSHIP ATTRIBUTION ON THE ENRON EMAIL CORPUS

A Thesis

Submitted to the McAnulty College & Graduate School of Liberal Arts

Duquesne University

In partial fulfillment of the requirements for

the degree of Masters of Science

By

Xuan Li

August 2013

AUTHORSHIP ATTRIBUTION ON THE ENRON EMAIL CORPUS

By

Xuan Li

Approved March 27, 2013

Patrick Juola Associate Professor, Department of Mathematics & Computer Science (Committee Chair)	Abhay Gaur, Ph.D. Professor, Department of Mathematics & Computer Science (Committee Member)
James Swindal, Ph.D., Dean, McAnulty College and Graduate School of Liberal Arts Professor of Philosophy	Donald Simon, Ph.D. Director of Graduate Study, Department of Mathematics & Computer Science

ABSTRACT

AUTHORSHIP ATTRIBUTION ON THE ENRON EMAIL CORPUS

By

Xuan Li

March, 2013

Thesis supervised by Patrick Juola

In this paper I present authorship attribution on an email corpus. The source I used was the Enron Email Corpus (Cohen, 2009). By reformatting these emails, four test sets were categorized based on the length of each email: Tiny ($\leq$ 99 characters), Small (100 to 500 characters), Medium (501 to 999 characters), and Large ($\geq$ 1000 characters). The Java Graphical Authorship Attribution Program (JGAAP software) from our Evaluating Variations in Language Laboratory (EVL Lab) was used to perform these tests. Three analysis methods: WEKA RandomForest, WEKA SMO, and Centroid with Cosine Distance were used. Results showed that the Large test set gave the best authorship classification, followed by the Medium, then the Small and the Tiny test sets. WEKA SMO gave better authorship classification than WEKA RandomForest.

DEDICATION

I sincerely dedicate this thesis to Professor Patrick Juola and the EVL Lab.

Without your enlightenment and guidance, I wouldn't be able to complete my thesis.

ACKNOWLEDGMENTS

My sincere thanks to:

Abhay Gaur,

Amanda Kroft,

Donald Simon,

Frank D'Amico,

John Kern,

Mike Ryan,

Pam Thompson,

Patrick Juola.

I really appreciate your help.

TABLE OF CONTENTS

LIST OF FIGURES

Page

Chapter 1. Introduction

Our Evaluating Variations in Language Laboratory (EVL Lab) has been conducting research in machine learning area, and we are working on authorship attribution. The modern principle behind authorship attribution is computer-based statistical measuring of textual features by different authors. In authorship attribution, textual documents are classified into two types, either with known authors or with unknown authors. In order to get correct authorship classifications on anonymous documents, the author who actually wrote the anonymous documents must be presented by some documents. By matching writing patterns and textual characteristics, the correct authorship then can be deduced. Without the author presented by documents, authorship would not be deduced correctly (e.g. given author C actually wrote the anonymous article, however, only author A and author B are presented by some articles. Then authorship attribution on that anonymous article can only be either author A or author B).

Our EVL Lab has done many tests on literatures such as novels, short stories, articles and even Tweets. We were wondering if we could get some authorship attribution on the most frequently used daily dialog: Email, which is the most popular communication tool of the current Internet. In this paper, I performed authorship attribution on an email corpus. My purpose was to see how accurate I could get authorship classifications on email.

Chapter 2. Background

2.1 Authorship Attribution

Authorship attribution can be defined as matching the most likely author with an anonymous textual document using existing examples of documents by the given authors.

Authorship attribution can be applied to plagiarism detection (e.g. papers or articles), analyzing the source of an unknown or allonymous textual document (e.g. threatening or harassing emails), and also classifying historical literature with unknown or unclear authorship (Bozkurt, Baghoglu, & Uyar, 2007). Authorship attribution is useful when there's a dispute about who has written the paper (either everyone says he or she has written it or no one is willing to admit he has written it). In authorship attribution, textual documents are classified into two types: documents with known authors (called training data), and documents with unknown authors (called testing data). By means of some specific computer-based statistical processing on the textual features, documents in the training data are mapped onto the multi-dimensional coordinate. Documents in the testing data also go through the statistical processing and get mapped onto the same coordinate. Through some computer-based statistical calculations and comparisons, documents in the testing data are matched with authors in the training data. Authorships of the testing data are classified.

2.2 JGAAP Software

The Java Graphical Authorship Attribution Program (JGAAP software) was developed by our EVL Lab to solve problems such as textual analysis, text categorization, and authorship attribution (Juola, 2007). The user interface of the JGAAP software is comprised of five parts. The first part: "Documents", where known and unknown documents are separately uploaded. The second part: "Canonicizers", which standardize all the documents. In this experiment, I used "None" (no Canonicizers), "Normalize Whitespace" (convert all whitespace characters to a single space), "Punctuation Separator" (put a single space before and after each punctuation mark), "Strip

Punctuation" (strip all punctuation characters), and "Normalize Whitespace" combined with "Punctuation Separator". The third part: "Event Drivers", where all the standardized documents are reformatted into subsets such as words. In this experiment I used "Words" (each subset is a word from the documents), "Character Grams 3" (each subset is 3 successive characters), and "Character Grams 4" (each subset is 4 successive characters). The fourth part: "Event Culling", where some specific choices are set according to the experimenter's preference (in this experiment, event culling was not used). The fifth part: "Analysis Methods", where the most widely adopted classification algorithms are incorporated. I used "Centroid Driver with Cosine Distance" (compute the distance of one centroid per author to another), "WEKA SMO" (Sequential Minimal Optimization (SMO) in the Java package form Waikato Environment for Knowledge Analysis (WEKA) which "is a workbench for machine learning that is intended to aid in the application of machine learning techniques to a variety of real-world problems" (Holmes, & Witten, 1994, p. 357; Juola, 2007)).

2.3 WEKA RandomForest

The EVL lab has been continually improving the JGAAP software to meet the latest requirements for authorship attribution. Breiman (2001) proposed a classification method: Random Forests which "are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest" (p. 5). Since the JGAAP software has the WEKA package already built-in, I incorporated the WEKA RandomForest classifier into JGAAP by referring to the Java class path: weka.classifiers.trees.randomforest. WEKA RandomForest (RF) has three parameters to set: "K –number of features to consider at

each split", "I –number of trees", and "S –number of seeds" with default settings "I = 10", "K = 0", "S = 1" (Hall et al., 2009). The mechanism behind WEKA RandomForest is: Draw *I* bootstrap samples from the original training data (documents with known authors). For each bootstrapped sample, randomly sample *K* features and choose the best split at each node to grow a classification tree (randomly sample $LogM+1$ when $K=0$, where *M* is the total number of features). Continue to work down the tree until no more nodes can be split. Predict new testing data (documents with unknown authors) by aggregating the predictions of the *I* trees (Liaw, & Wiener, 2002). Due to the random property of WEKA RandomForest, setting "S=1" allows that this experiment can be reproduced. In order to get more classification trees in my experiment, I set default "I" to "1000".

2.4 The Enron Email Corpus

Our EVL lab has done many authorship attribution tests on different kinds of works using JGAAP software, and the results have been very good. We are wondering if JGAAP can also work on short conversations such as emails. We turned to the Enron Email Corpus which is appealing to researchers because it is a real large-scale corporate email collection. The Enron Corporation was once the seventh largest business organization in the USA. However, in 2001 the organization announced itself bankrupt. A corpus of emails from the Enron Corporation was made public during the legal investigation by the Federal Energy Regulatory Commission (FERC) (Diesner, Frantz, & Carley, 2005; Klimt, & Yang, 2004). The raw email dataset contains more than 500,000 messages from about 150 senior management executives at the Enron Corporation (Shetty, & Adibi,

2004; Styler, 2011). William Cohen (2009) of Carnegie-Mellon University has put up the Enron dataset on the web free for research use (http://www.cs.cmu.edu/~enron/).

Chapter 3. Materials and Methods

3.1 Email Process

The Enron Email Corpus contains about 150 authors. All the emails are organized into folders. The difficulty in using the Enron corpus is these emails had not been formalized or reformatted. It includes all kinds of emails, such as spam and computer-generated messages. Since a substantial portion of the emails is non-human-written, a reasonable approach is to only look at the "sent" folders. Even though emails from one "sent" folder are mostly written by that author, there are still considerably large pieces of "useless" text: forwarded, replied, and other computer-generated messages such as headings. To extract just the contents from emails in the "sent" folders and save them into text files, I wrote Java code to perform the task. The Java program read each message from the "sent" folder. It bypassed the email header until the body which is characterized by one space lines, and it saved all the text until meeting any one of the ending signals: three space lines, a line started with "------ forwarded", a line started with "------ replied", or the end of the message. According to the length of the extracted emails, I categorized them into four test sets: Tiny ($\leq$ 99 characters), Small (100 to 499 characters), Medium (500 to 999 characters), and Large ($\geq$ 1000 characters). Based on how many reformatted messages were left, 5 to 12 emails were randomly selected from each author for each test set. When selecting emails, I also manually deleted human name entities appearing in greetings (e.g. "Hey John" will be "Hey ") and signatures, as I was concerned of these names might render inaccurate classifications. When I was deleting name entities, I

noticed some emails from authors (most likely high position executives at the Enron Corporation) were written by their secretaries (e.g. words "on behalf of" were used in the signature). In case of confusion, I deleted all emails from those authors.

3.2 Classification Methods

I tested the following Canonicizers: None, Punctuation Separator (PS), Strip Punctuation (SP), Normalize Whitespace (NW), and Normalize Whitespace combined with Punctuation Separator, and the following Event Drivers: Words, Character Grams 3 and 4. The Analysis Methods I used were: WEKA Sequential Minimal Optimization (SMO) Classifier, Centroid Driver plus Cosine Distance (Centroid approach needs to work on a given distance function), and WEKA RandomForest (RF) Classifier. WEKA SMO and Centroid classifiers had previously been proven to be effective on authorship attribution. Even with the default settings of WEKA SMO, one can get quite accurate classifications. I tested the newly incorporated WEKA RF classifier, and compared the results with other classification methods. To have an unbiased comparison, both WEKA SMO and WEKA RandomForest were used with default values (Centroid Driver does not have parameters). Parameters for Canonicizers, Event Drivers, and Analysis Methods are shown in table 1. 10-fold crossvalidation was used to perform the experiment. The Enron Email Corpus was randomly divided into 10 sections (Computer-generated randomization). Each time, one section was set as testing data (unknown document), and the rest nine sections were set as training data (known document), until every group set as testing data once. A Java program was written to count the authors and their emails to ensure that when doing the 10-fold crossvalidation, not all the emails of an author's went into the same section

(Authorship attribution requires that an author has at least one textual document in the training data)

Table 2. JGAAP Parameter Settings

Parameter combination index	Canoniciers	Event Drivers	Analysis Methods
1	None	Character Grams 3	Centroid|Cosine Distance
2	None	Character Grams 3	WEKA RF
3	None	Character Grams 3	WEKA SMO
4	None	Character Grams 4	Centroid|Cosine Distance
5	None	Character Grams 4	WEKA RF
6	None	Character Grams 4	WEKA SMO
7	None	Words	Centroid|Cosine Distance
8	None	Words	WEKA RF
9	None	Words	WEKA SMO
10	Normalize Whitespace (NW)	Character Grams 3	Centroid|Cosine Distance
11	Normalize Whitespace	Character Grams 3	WEKA RF
12	Normalize Whitespace	Character Grams 3	WEKA SMO
13	Normalize Whitespace	Character Grams 4	Centroid|Cosine Distance
14	Normalize Whitespace	Character Grams 4	WEKA RF
15	Normalize Whitespace	Character Grams 4	WEKA SMO
16	NW|PS	Character Grams 3	Centroid|Cosine Distance
17	NW|PS	Character Grams 3	WEKA RF
18	NW|PS	Character Grams 3	WEKA SMO
19	NW|PS	Character Grams 4	Centroid|Cosine Distance
20	NW|PS	Character Grams 4	WEKA RF
21	NW|PS	Character Grams 4	WEKA SMO
22	Punctuation Separator (PS)	Character Grams 3	Centroid|Cosine Distance
23	Punctuation Separator	Character Grams 3	WEKA RF
24	Punctuation Separator	Character Grams 3	WEKA SMO
25	Punctuation Separator	Character Grams 4	Centroid|Cosine Distance
26	Punctuation Separator	Character Grams 4	WEKA RF
27	Punctuation Separator	Character Grams 4	WEKA SMO
28	Punctuation Separator	Words	Centroid|Cosine Distance
29	Punctuation Separator	Words	WEKA RF
30	Punctuation Separator	Words	WEKA SMO
31	Strip Punctuation (SP)	Character Grams 3	Centroid|Cosine Distance
32	Strip Punctuation	Character Grams 3	WEKA RF
33	Strip Punctuation	Character Grams 3	WEKA SMO
34	Strip Punctuation	Character Grams 4	Centroid|Cosine Distance
35	Strip Punctuation	Character Grams 4	WEKA RF
36	Strip Punctuation	Character Grams 4	WEKA SMO
37	Strip Punctuation	Words	Centroid|Cosine Distance
38	Strip Punctuation	Words	WEKA RF
39	Strip Punctuation	Words	WEKA SMO

Chapter 4. Results

In the Large set, there were a total of 368 emails with 36 authors. In the Medium set, there were a total of 614 emails with 56 authors. In the Small set, there were a total of 822 emails with 69 authors, and in the Tiny set, there were a total of 842 emails with 70

authors. The percentages of correct classifications: accuracy (correct count divided by overall count) according to the parameter combinations are shown in table 2.

Table 2. Accuracy by set

Parameter combination index	Large set	Medium set	Small set	Tiny set
1	0.4592	0.3094	0.2433	0.1461
2	0.3696	0.2622	0.2190	0.1354
3	0.5353	0.3762	0.2479	0.1366
4	0.4674	0.3111	0.2384	0.1603
5	0.3560	0.2801	0.1691	0.1283
6	0.4946	0.3143	0.2141	0.1045
7	0.3668	0.1840	0.1314	0.1069
8	0.4701	0.3127	0.2141	0.1354
9	0.4375	0.2932	0.1861	0.1093
10	0.4429	0.2638	0.1934	0.1223
11	0.3886	0.2508	0.1533	0.1223
12	0.5272	0.3550	0.2044	0.1152
13	0.4620	0.2818	0.2032	0.1223
14	0.3587	0.2687	0.1557	0.1081
15	0.5027	0.2883	0.1971	0.0914
16	0.4837	0.3094	0.2129	0.1342
17	0.3832	0.2345	0.1837	0.1271
18	0.5190	0.3502	0.2129	0.1200
19	0.4837	0.3290	0.2214	0.1366
20	0.3587	0.2671	0.1752	0.1176
21	0.5163	0.3127	0.2129	0.1140
22	0.5109	0.3274	0.2628	0.1580
23	0.3750	0.2785	0.2178	0.1461
24	0.5353	0.3876	0.2470	0.1508
25	0.5082	0.3550	0.2689	0.1817
26	0.3913	0.2883	0.2178	0.1306
27	0.5082	0.3274	0.2251	0.1211
28	0.4647	0.2427	0.1788	0.1318
29	0.4837	0.3257	0.2105	0.1485
30	0.5082	0.3518	0.2445	0.1366
31	0.4402	0.2524	0.1776	0.1116
32	0.3587	0.2280	0.1703	0.1105
33	0.5109	0.3534	0.1983	0.1152
34	0.4484	0.2671	0.2032	0.1152
35	0.3533	0.2524	0.1582	0.1093
36	0.4620	0.2801	0.1727	0.0950
37	0.3967	0.1906	0.1168	0.0784
38	0.4457	0.2980	0.1800	0.0998
39	0.4701	0.3046	0.1776	0.0926
Mean	0.4501	0.2940	0.2004	0.1238
Std Dev	0.0594	0.0467	0.0343	0.0210
Median	0.4620	0.2932	0.2032	0.1223

Chapter 5. Computations

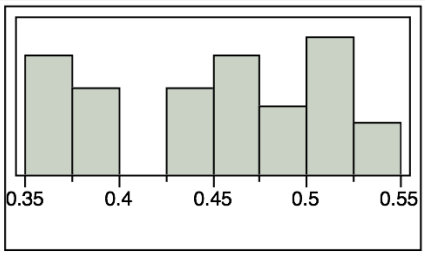
5.1 Test by Set

JMP software was used for statistical computing. Accuracy distributions by set are shown from Fig 1 to Fig 4.

Fig 1. Large set distribution

Distributions Size=Large

Accuracy



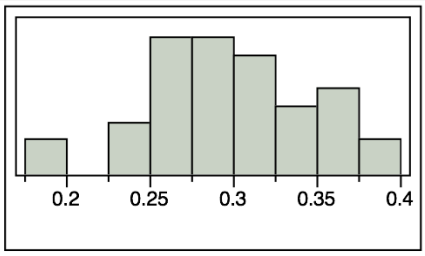
Quantiles		
100.0%	maximum	0.5353
99.5%		0.5353
97.5%		0.5353
90.0%		0.519
75.0%	quartile	0.5082
50.0%	median	0.462
25.0%	quartile	0.3886
10.0%		0.3587
2.5%		0.3533
0.5%		0.3533
0.0%	minimum	0.3533

Moments	
Mean	0.4501205
Std Dev	0.0594201
Std Err Mean	0.0095148
Upper 95% Mean	0.4693823
Lower 95% Mean	0.4308587
N	39

Fig 2. Medium set distribution

Distributions Size=Medium

Accuracy



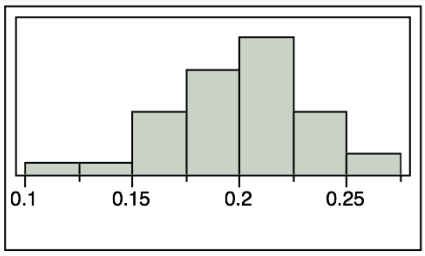
Quantiles		
100.0%	maximum	0.3876
99.5%		0.3876
97.5%		0.3876
90.0%		0.355
75.0%	quartile	0.3274
50.0%	median	0.2932
25.0%	quartile	0.2638
10.0%		0.2345
2.5%		0.184
0.5%		0.184
0.0%	minimum	0.184

Moments	
Mean	0.2939872
Std Dev	0.0467225
Std Err Mean	0.0074816
Upper 95% Mean	0.3091329
Lower 95% Mean	0.2788415
N	39

Fig 3. Small set distribution

Distributions Size=Small

Accuracy



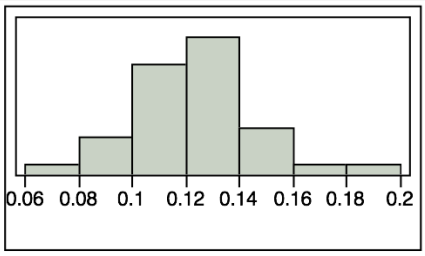
Quantiles		
100.0%	maximum	0.2689
99.5%		0.2689
97.5%		0.2689
90.0%		0.247
75.0%	quartile	0.219
50.0%	median	0.2032
25.0%	quartile	0.1776
10.0%		0.1557
2.5%		0.1168
0.5%		0.1168
0.0%	minimum	0.1168

Moments	
Mean	0.2004462
Std Dev	0.0343085
Std Err Mean	0.0054938
Upper 95% Mean	0.2115677
Lower 95% Mean	0.1893246
N	39

Fig 4. Tiny set distribution

Distributions Size=Tiny

Accuracy



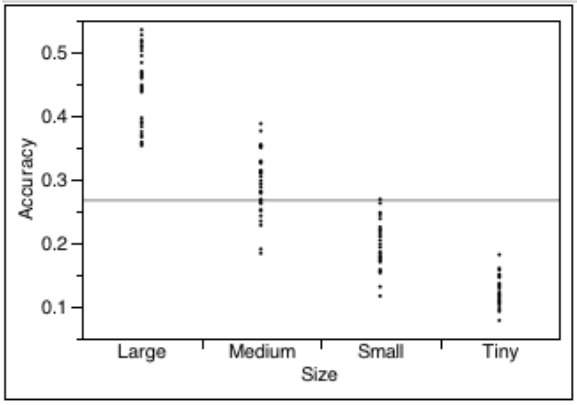
Quantiles		
100.0%	maximum	0.1817
99.5%		0.1817
97.5%		0.1817
90.0%		0.1508
75.0%	quartile	0.1366
50.0%	median	0.1223
25.0%	quartile	0.1093
10.0%		0.095
2.5%		0.0784
0.5%		0.0784
0.0%	minimum	0.0784

Moments	
Mean	0.1237615
Std Dev	0.0209784
Std Err Mean	0.0033592
Upper 95% Mean	0.130562
Lower 95% Mean	0.1169611
N	39

Since data did not come from normal distributions, the Nonparametric Wilcoxon Test which is equivalent to Kruskal-Wallis test when more than two groups (Kruskal, & Wallis, 1952) and the Nonparametric Multiple Comparisons in JMP were used with significance level at 0.005 (adjusted α-value for multiple tests). Both tests used medians and ranks of the accuracies for statistical computations.

Fig 5. Nonparametric test by set

One-way Analysis of Accuracy By Size



Wilcoxon / Kruskal-Wallis Tests (Rank Sums)

Level	Count	Score Sum	Expected Score	Score Mean	(Mean-Mean0)/Std0
Large	39	5323.00	3061.50	136.487	9.254
Medium	39	3761.00	3061.50	96.436	2.861
Small	39	2329.00	3061.50	59.718	-2.996
Tiny	39	833.000	3061.50	21.359	-9.119

1-way Test, ChiSquare Approximation

ChiSquare	DF	Prob>ChiSq
139.5404	3	<.0001*

Nonparametric Comparisons For Each Pair Using Wilcoxon Method

q*	Alpha
2.80703	0.005

Level	- Level	Score Mean Difference	Std Err Dif	Z	p-Value	Hodges-Lehmann	Lower CL	Upper CL
Small	Medium	-34.8205	5.131082	-6.78619	<.0001*	-0.094400	-0.121900	-0.066000
Tiny	Small	-36.2564	5.130855	-7.06635	<.0001*	-0.077500	-0.096800	-0.058400
Medium	Large	-37.9487	5.130823	-7.39622	<.0001*	-0.159100	-0.197600	-0.117200
Small	Large	-38.9744	5.130823	-7.59612	<.0001*	-0.257600	-0.290100	-0.216300
Tiny	Large	-38.9744	5.130596	-7.59646	<.0001*	-0.336800	-0.366100	-0.291700
Tiny	Medium	-38.9744	5.130855	-7.59607	<.0001*	-0.170900	-0.195400	-0.146400

As shown above, Kruskal-Wallis Test was significant at 0.005 (α-value), which meant that the median accuracies from all four test sets were not equal. Further Nonparametric Multiple Comparisons confirmed that each paired test was significant at 0.005 level. According to the confidence interval in the Multiple Comparisons, we are 99.5% certain

that [-0.122, -0.066] contains the true difference of Accuracy(Small) – Accuracy(Medium); [-0.097, -0.058] contains the true difference of Accuracy(Tiny) – Accuracy(Small); [-0.198, -0.118] contains the true difference of Accuracy(Medium) – Accuracy(Large); [-0.290, -0.216] contains the true difference of Accuracy(Small) – Accuracy(Large); [-0.366, -0.292] contains the true difference of Accuracy(Tiny) – Accuracy(Large); [-0.195, -0.146] contains the true difference of Accuracy(Tiny) – Accuracy(Medium). From above, authorship attribution on the Enron Email Corpus shows: Accuracy(Large) > Accuracy(Medium) > Accuracy(Small) > Accuracy(Tiny).

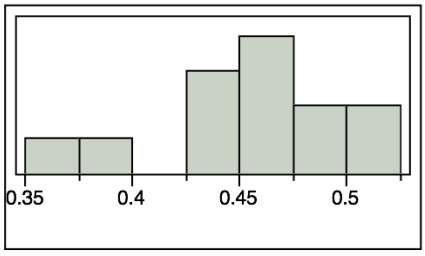
5.2 Test by Analysis Method

In authorship attribution, we would want accuracies as high as possible. Therefore, it makes sense to study the data in the Large set, since the three analysis methods may generate significantly different accuracies.

Fig 6. Distribution by Centroid

Distributions Analysis Methods=Centroid

Accuracy in Large



Quantiles

100.0%	maximum	0.5109
99.5%		0.5109
97.5%		0.5109
90.0%		0.50982
75.0%	quartile	0.4837
50.0%	median	0.462
25.0%	quartile	0.44155
10.0%		0.37876
2.5%		0.3668
0.5%		0.3668
0.0%	minimum	0.3668

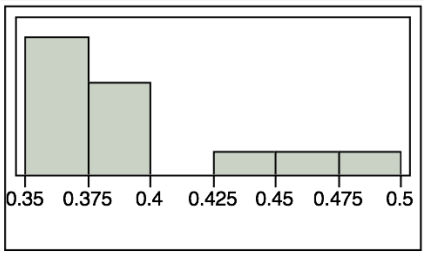
Moments

Mean	0.4565231
Std Dev	0.0402843
Std Err Mean	0.0111729
Upper 95% Mean	0.4808666
Lower 95% Mean	0.4321795
N	13

Fig 7. Distribution by WEKA RandomForest

Distributions Analysis Methods=RF

Accuracy in Large



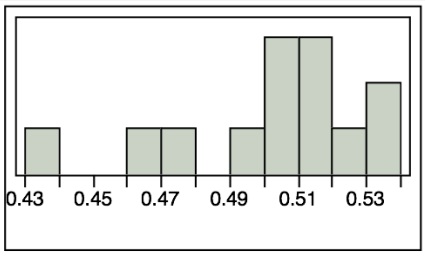
Quantiles		
100.0%	maximum	0.4837
99.5%		0.4837
97.5%		0.4837
90.0%		0.47826
75.0%	quartile	0.4185
50.0%	median	0.375
25.0%	quartile	0.3587
10.0%		0.35438
2.5%		0.3533
0.5%		0.3533
0.0%	minimum	0.3533

Moments	
Mean	0.3917385
Std Dev	0.045085
Std Err Mean	0.0125043
Upper 95% Mean	0.4189831
Lower 95% Mean	0.3644939
N	13

Fig 8. Distribution by WEKA SMO

Distributions Analysis Methods=SMO

Accuracy in Large



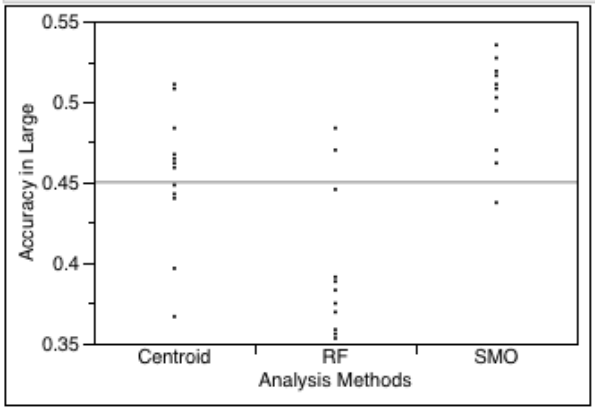
Quantiles		
100.0%	maximum	0.5353
99.5%		0.5353
97.5%		0.5353
90.0%		0.5353
75.0%	quartile	0.5231
50.0%	median	0.5082
25.0%	quartile	0.48235
10.0%		0.4473
2.5%		0.4375
0.5%		0.4375
0.0%	minimum	0.4375

Moments	
Mean	0.5021
Std Dev	0.0293476
Std Err Mean	0.0081396
Upper 95% Mean	0.5198346
Lower 95% Mean	0.4843654
N	13

Data did not come from normal distributions. Nonparametric Wilcoxon Test and Nonparametric Multiple Comparisons in JMP were used with significance level at 0.005. Medians and ranks were used for statistical computations.

Fig 9. Nonparametric test by analysis method

One-way Analysis of Accuracy in Large By Analysis Methods



Excluded Rows 117

Wilcoxon / Kruskal-Wallis Tests (Rank Sums)

Level	Count	Score Sum	Expected Score	Score Mean	(Mean-Mean0)/Std0
Centroid	13	261.000	260.000	20.0769	0.015
RF	13	125.500	260.000	9.6538	-3.995
SMO	13	393.500	260.000	30.2692	3.966

1-way Test, ChiSquare Approximation

ChiSquare	DF	Prob>ChiSq
21.2851	2	<.0001*

Nonparametric Comparisons For Each Pair Using Wilcoxon Method

q*	Alpha
2.80703	0.005

Level	- Level	Score Mean Difference	Std Err Dif	Z	p-Value	Hodges-Lehmann	Lower CL	Upper CL
SMO	RF	11.9231	2.996408	3.97912	<.0001*	0.125000	0.051600	0.1576000
SMO	Centroid	8.4615	2.995895	2.82438	0.0047*	0.046200	0.000000	0.0869000
RF	Centroid	-8.6154	2.995895	-2.87573	0.0040*	-0.081500	-0.119600	0.0000000

Kruskal-Wallis Test was significant at 0.005 (α-value), which meant that the median accuracies from the three analysis methods were not equal. Nonparametric Multiple Comparisons showed that each paired test was significant at 0.005 level. However, according to the confidence interval, we are 99.5% certain that [0.052, 0.158] contains the true difference of Accuracy(SMO) – Accuracy(RandomForest); [0.000, 0.087]

contains the true difference of Accuracy(SMO) – Accuracy(Centroid); [-0.120, 0.000] contains the true difference of Accuracy(RandomForest) – Accuracy(Centroid). From the above confidence intervals, we can say Accuracy(SMO) > Accuracy(RandomForest). However, there are only slight differences between Accuracy(SMO) and Accuracy(Centroid), and between Accuracy(Centroid) and Accuracy(RandomForest).

Chapter 6. Discussions and Suggestions

From the above results, it showed that emails in the Large set ($\geq$ 1000 characters) generated the best authorship attribution with a median accuracy of 46.2%. Emails in the Medium set (500 to 999 characters) rendered the second authorship attribution with a median accuracy of 29.3%. Emails in the Small set (100 to 499 characters) rendered accuracy with a median of 20.3%, and emails in the Tiny set ($\leq$ 99 characters) with a median of 12.2%. The results demonstrated that the larger the emails, the better accuracy on authorship attribution. This experiment also revealed some differences in efficacies from different analysis methods. In the Large set, WEKA SMO gave a range of accuracies with a median of 50.8%. Centroid Driver with Cosine Distance gave a range of accuracies with a median of 46.2%, and WEKA RandomForest with a median accuracy of 37.5% (One thing needs to be noted: when running tests on JGAAP, WEKA RandomForest Classifier took much longer time than the other two). Statistical computations confirmed that WEKA SMO gave better authorship attribution than WEKA RandomForest. However, the differences between WEKA SMO and Centroid, and between Centroid and WEKA RandomForest were slight.

This experiment showed that authorship attribution on emails would require email length larger than 500 characters (the Small set (100 to 499 characters) gave an accuracy around

20% and is of little worth). Even though WEKA SMO is slightly better than Centroid Driver, WEKA SMO was with the default settings. If we want to increase the accuracy on email authorship attribution, one feasible way would be to optimize parameters of WEKA SMO. Based on the results, WEKA SMO seemed to work best with the Event Driver of Character Gram 3, and both WEKA SMO and Centroid Driver worked worst with Words. It might be useful to test WEKA SMO and Centroid Driver with Character Gram 2, respectively. Even though I manually deleted human name entities appearing in greeting and signature, the email contents still have some name entities left. Some names frequently appeared in the corpus (e.g. David, John, and Davis, etc.). Since the Enron Email Corpus is from a real corporate email collection, written by 150 employees at the Enron Corporation, it is common that some executives' names were frequently mentioned. However, these name entities might affect the accuracy of authorship attribution. To improve the accuracy, our EVL Lab should be working to program a Canonicizer to detect and delete name entities appearing in the text.

REFERENCES

Bozkurt, I. N., Baghoglu, O., & Uyar, E. (2007, November). Authorship attribution. In *Computer and information sciences, 2007. iscis 2007. 22nd international symposium on* (pp. 1-5). doi: 10.1109/ISCIS.2007.4456854

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Cohen, W. W. (2009, August). *Enron email dataset.* Retrieved from Carnegie Mellon University website: http://www.cs.cmu.edu/~enron/

Diesner, J., Frantz, T. L., & Carley, K. M. (2005). Communication networks from the Enron email corpus "It's always about the people. Enron is no different". *Computational & Mathematical Organization Theory*, *11*(3), 201-228.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *SIGKDD Explorations*, *11*(1).

Holmes, G., Donkin, A., & Witten, I. H. (1994, December). WEKA: A machine learning workbench. In *Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference* (pp. 357-361). IEEE.

Juola, P. (2007). Authorship attribution. *Foundations and Trends® in Information Retrieval*, *1*(3), 233-334.

Klimt, B., & Yang, Y. (2004). The Enron corpus: A new dataset for email classification research. *Machine Learning: ECML 2004*, 217-226.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*(260), 583-621.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R news*, *2*(3), 18-22.

Shetty, J., & Adibi, J. (2004). The Enron email dataset database schema and brief statistical report. *Information Sciences Institute Technical Report, University of Southern California*, *4*.

Styler, W. (2011, May). *The EnronSent corpus* [PDF format]. Retrieved from University of Colorado Boulder website: http://www.colorado.edu/ics/sites/default/files/attached-files/01-11_0.pdf